Differences in Breast Cancer Sub-types

Emma Spors — emma.spors@jacks.sdstate.edu
Department of Mathematics and Statistics
South Dakota State University

Fall 2021

Abstract

Motivation: Breast cancer (BC) affects 1 in 8 women during their lifetime. This makes it the second most common cancer and second leading cause of cancer death in women. However, not all breast cancers are "created" equal, with some more aggressive sub-types being associated with a worse prognosis than others. Thus, with advancements in bioinformatics, the genetic differences between BC sub-types, such as luminal and basal, has become of huge interest to the cancer community as a prognostic and therapeutic tool. One such study has examined the gene Wwox and it's role in the JAK/STAT pathway. This analysis strove to recreate their results to better understand the roles of Wwox and the JAK/STAT pathway in breast cancer.

Data and Methodology: The data from the previous study was made publicly available on the Gene Expression Omnibus repository. It consisted of six libraries in total from three luminal breast cancer cell lines and three basal breast cancer cell lines. Differential gene expression was completed with the R package *DESeq2* with absolute log fold change greater than 2 and adjusted p-value less than 0.05. Pathway analysis was completed with the R package *gage* using Go Ontology and KEGG annotated pathways.

Results: Around 1,500 genes were differently expressed between the basal and luminal breast cancers. However, the gene of interest Wwox was not significantly down-regulated in the basal cells compared to the luminal. The JAK/STAT pathway was up-regulated in the basal BCs which was consistent with the paper's results.

Discussion: While this analysis did not find the Wwox gene to be significantly down-regulated, some studies suggest that the relative expression of this gene differs widely in cell lines. In addition, other data sources are congruent with the findings of the original study. Up-regulation of the JAK/STAT pathway was consistent with other studies, stating the STAT3 gene in the pathway is know to cause malignant behavior in cancers. Ultimately, these two results may be candidates for further study.

Differences in Breast Cancer Sub-types

A short analysis of "Loss of Wwox drives metastasis in triple-negative breast cancer by JAK2/STAT3 axis" by Chang et al.

1 Background and Motivation

Around 1 in 8 women will develop breast cancer (BC) during her lifetime. This makes it the second most common cancer in women and the second leading cause of cancer death [1]. Still, with scientific advances, BC-related deaths have been slowly decreasing since the nineties. However, not all breast cancers are "created" equal with some more aggressive sub-types being associated with a worse prognosis. Thus, with the advancements in bioinformatics and molecular profiling, the genetic differences between BC sub-types have become a huge interest to the scientific community for prognostic and therapeutic applications.

1.1 Breast Cancer Sub-types

There are many different kinds of breast cancer which most commonly include ductal cacinomca in situ (noninvasive or pre-invasive) and invasive (cancer that has spread into the surrounding breast tissue) . Approximately 70-80 percent of all BCs are invasive [2]. In addition, there are four to five molecular sub-types of BC. Luminal A is BC that is hormone-receptor positive, HER2 negative, and has low levels of protein Ki-67. The hormone receptors refer to estrogen-receptor and progesterone-receptor. HER2 is the human epidermal growth factor receptor 2, a protein found on the surface of breast cells. The protein Ki-67 helps to measure a cancer's aggressiveness. Luminal B is BC that is hormone-receptor positive, either HER2 negative or positive, with high levels of Ki-67. Normal-like, which is sometimes combined with Luminal A, is hormone-receptor positive, HER2 negative, and has low levels of Ki-67. HER2-enriched BC is hormone receptor negative and HER2 positive [3].

Triple-negative is BC that is hormone receptor negative and HER2 negative. Triple-negative breast cancer (TNBC) is sometimes used interchangeably with basal-like breast cancer as they are both similar, aggressive forms of breast cancer. However, TNBC refers specifically to the aforementioned traits, making it a narrow classification of tumor. Basal-like tumors refer to a much wider class of tumors that all have "basal-like" cells. Still, up to 70 percent of TNBC are basal and 77 percent of basal BC are TNBC. Of these sub-types, triple-negative breast cancer (TNBC) typically has the worst prognosis as it is more likely to metastasize and recur after treatment. In addition, since the growth is not fueled by the hormone estrogen or progesterone, or by the HER2 protein, TNBC does not respond to current hormonal or HER2 tageted therapies [3]. This paper will focus on the luminal and basal sub-types.

1.2 Literature Review

This study followed the paper "Loss of Wwox drive metastasis in triple-negative breast cancer by JAK2/STAT3 axis" by Change et al. which compared TNBC/basal-like breast cancer to luminal breast cancer [4]. The researchers performed gene expression analysis to determine that Wwox was down-regulated in basal cells (as compared to luminal) and that it's lower expression was significantly correlated with the activation of gene STAT3. Consequently, the JAK/STAT

pathway was one of the most differently modulated pathways between the sub-types. In addition, they performed in vivo and in vitro experiments to demonstrate that over-expression of Wwox in BC cells inhibited the proliferation and metastasis of the cells by suppressing the STAT3 activity. Cnonversely, the knock-down of Wwox caused metastatic-like behaviors. They concluded that the targeting of STAT3 through Wwox offers a potential therapeutic tool for TNBC.

1.3 Wwox

WW domain-containg oxidoreductase (Wwox) is a gene that codes for the enzyme of the same name. The gene is located on chromose 16 and is encoded by a locus that is one of the most common fragile sites involved with cancer because it is highly susceptible to DNA damage [5]. Despite its precarious location, some evidence suggests that Wwox acts a tumor suppressor and thus has become a protein of interest in recent years. Wwox's other duties include interacting with transcription regulators to mediate the action of cytokines, interferons, and growth hormone [5].

1.4 STAT3 and JAK/STAT Pathway

Signal transducer and activator of transcription 3 (STAT3) is a member of the broader STAT protein family which act as transcription factors. This protein regulates other genes that control such important cellular processes as proliferation, survival, and motility [6]. Typically STAT3 is tightly regulated, but during some strenuous event, STAT3 can become highly activated which causes the elevation of downstream genes which drives malignant cell behavior. However, other studies have found that the up-regulation STAT3 is not synonymous with a cancerous outcome but may rather act as a tumor suppressor [6]. Thus, additional study on STAT3 is needed.

The janus kinase and signal transducer and activator of transcription (JAK/STAT) pathway, of which STAT3 is a part, is a signaling pathway regulated the JAK and STAT proteins [7]. Since it's discovery several decades ago, this pathway has quickly become of the best studied intra-cellular signaling networks. The pathway helps to control apoptosis, cell cycle progression/inhibition, differentiation, proliferation, and cell survival. Deregulation of this pathway is associated with several diseases, including immune system disorders and cancers [7].

1.5 Objective

Thus motivated by the outcomes of the study by Chang et al. and the uncertainty surrounding the gene/protein Wwox and the JAK/STAT signaling pathway, this analysis sought to recreate the gene expression and pathway analysis previously conducted to compare results and gain new understandings.

2 Data and Methodology

2.1 Data

The gene expression data from the study was made publicly available on the Gene Expression Omnibus repository under the accession code GSE110810. There were 6 libraries in total consisting of three luminal breast cancer lines (MCF-7, T47D, and BT-474) and 3 basal breast cancer cells (SUM-159, HBL-100, BT-549). Both raw read counts and FPKM were available, but only the counts were downloaded for this analysis. It is worth noting that the HBL-100 cell line

was not derived from cancerous cells, rather is a spontaneously immortalized cell line that came from the breast milk of a young, healthy, lactating woman. However, HBL-100 does not express hormone-receptors or have amplified HER2 and the DNA has transforming activity in vitro [8].

2.2 Methodology

Analysis was conducted with the software R and RStudio [9, 10]. The package *DESeq2* was used for differential gene expression analysis and the package *gage* was used for pathway analysis [11, 12]. General data wrangling and visualization was done with Tidyverse [13]. For code availability, please see the GitHub repository linked here.

Data Preparation. Data preparation was minimal. There were two missing values in all libraries. Because of the low number of missing values, these were set to zero. The experimental design only accounts for the sub-types. After the *DESeq2* object was created, the genes were filtered to be at least 0.5 counts per million in at least three of the libraries. This was necessary as very low gene counts can cause problems and lead to unreliable results. This reduced the number of genes from 50,869 to 17,057.

Exploratory Data Analysis. Next, exploratory data analysis (EDA) was performed. A variance stabilizing transformation (VST) was used to normalize the data for visualization and account for standard deviation's dependence on the mean. Methods included hierarchical clustering and PCA analysis. These were used to see similarities and differences in the data as well as check for irregularities.

Differential Gene Expression. *DESeq2* internally corrects for differences in library size, which is why it is important that we supplied the un-normalized counts of the sequencing reads. *DESeq2* models the read counts as following a negative binomial distribution where the mean is taken to be proportional to the concentration of cDNA fragments from the gene in the sample which is scaled by a normalization factor. In most cases, the scaling factor is the same for all genes in a samples which what accounts for difference in sequencing depth between the libraries [11].

In the style of the original paper, the adjusted p-value was set to 0.05 and the log fold change (LFC) was 0 as parameters. In other words, the null hypothesis was that the gene expression levels between basal and luminal was the same and the alternative hypothesis was that they were different. The adjusted p-value measures the probability the types are different by chance and is adjusted for multiple comparisons.

Pathway Analysis. Pathway analysis was conducted with all LFCs, regardless of significance, with *gage* methodology using KEGG and Gene Ontology pathways as a database [12]. GAGE, or generally applicable gene-set enrichment for pathway analysis, identifies biological pathways that are enriched more than what would be expected by chance. Using all LFCs, GAGE assumes that the gene set comes from a different distribution than the background set and thus uses a two-sample t-test to both account for gene set specific variance & background variance and to compare the expression level changes. Moreover, GAGE assumes that mean LFCs are normally distributed and are are independent. However, research has shown that these conditions are almost always met [12].

3 Results

3.1 Exploratory Data Analysis

Exploratory data analysis was used to gain insights into the data. For data visualization, we used a VST transformation which was confirmed to have remove most of the dependence of the standard deviation on the mean, see Section 6, Figure 6. This normalized the data, so genes with larger counts do not dominate the results in EDA.

PCA revealed that a large majority of the variation (66 percent) in the data was explained by difference between sub-types, Figure 1. Interestingly, the second component of PCA demonstrated the variation between the basal samples, indicating that they were somewhat dis-similar to each other. The hierarchical clustering also clustered the sub-types together 2. The top 30 genes were clustered by those that tended to have higher values in the basal libraries and those that had lower values in the basal libraries. This made for clear separations between the sub-types. Thus, we had some intuition that there would be a large number of differently expressed genes between the sub-types. There were no irregularities in the data noted at this stage.

3.2 Differently Expressed Genes

After applying the filters of an absolute value of LFC greater or equal to two and an adjusted p-value of less than 0.05, there were 793 up-regulated genes and 783 down-regulated genes. Thus, about nine percent of the filtered genes were differently expressed. A volcano plot of the results can be seen in Figure 4. However, the gene of interest, Wwox, was not significant. It did have a LFC of -2.1 (suggesting it was down-regulated in basal cells), but the adjusted p-value was 0.1. By looking at a plot of the normalized counts, Figure 5, we saw that the basal sub-types did have fewer counts of the gene on average. However, this is mainly due to one luminal library, MCF-7. In comparison, the genes secreted protein acid rich in cysteine (SPARC) and moesin had very high LFC (12.1 and 11.9) and low adjusted p-values ($3.3e^{-61}$ and $2.6e^{-62}$), suggesting that they were up-regulated in the basal cells. Indeed, research has suggested that both genes show prognostic significance for triple-negative breast cancer [14, 15]. Conversely, DSCAM-AS1 had a negative LFC of -13.8 and adjusted p-value of $2.2e^{-15}$, suggesting it was down-regulated in the basal cells. Studies have shown that this gene is dis-regulated in breast cancer which leads to proliferation and migration, so it is interesting that it was expressed in higher levels in the luminal cancers [16]. Additional research would be needed into this gene.

3.3 Pathway Analysis

Pathway analysis was completed using GO and KEGG Pathways. The results can be seen in Table 1 and Table 2. Of the top five GO pathways, ranked by p-value, all were from the biological process pathways and were up-regulated in the basal cells. The top four pathways, vasculature development, blood vessel development, extracellular matrix organization, and blood morphogenesis, all refer to the development or maintenance of physical structures in an organism necessary for formation of new structures, particularly for blood vessels.

For the KEGG pathways, the top-five most significant pathways were also up-regulated. First, focal adhesion relates to the cell matrix adhesion's necessary for cell proliferation, cell differentiation, and regulation of gene express for cell survival. ECM-receptor interaction relates to the actually processing of information that happens from environmental stimuli. Complement and

coagulation cascades pathway is involved with the immune system and protein digestion and absorption is quite literally the pathway for the digestion of proteins. All of these pathways might be expected to be up-regulated in basal BC. For the focus of this paper, we have the JAK-STAT signaling pathway as moderately significant. The accompanying map for the pathway can be seen in Figure 3.

4 Discussion

It is clear that there are many differently expressed genes between the basal and luminal breast cancer sub-types as 1,576 genes were found to be significant. Some we found, such as moesin and SPARC which were up-regulated in basal breast cancer cells, have literature that agree with our findings [14, 16]. DSCAM-AS1, which was down-regulated is associated with breast cancer, but it's role between sub-types is less clear. However, the main focus for this analysis was the Wwox gene.

The Wwox gene was not found to be significant during the differential expression analysis. The LFC and visual inspection of the counts do suggest that the basal BC may have had lower expression consistent with the original study, but the adjusted p-value was relatively high at 0.1. However, this may not be the end of the road for this gene. One study found that breast cancer cell lines range considerably in expression levels from undetectable to very high, and the expression was negatively correlated to the age of the individual they were taken from [5]. Thus, a part of the issue may be from the source material. Indeed, the study by Change et al. also found lower levels of Wwox in basal cancer cells from sequencing tissues samples not just cell lines and other studies in a clinical setting from have found similar results between cancers that were estrogen receptor (ER) positive and ER negative [4, 5]. Thus, the overall consensus is that Wwox does differently express between sub-types, but further research is needed to understand it's nuances.

Using all log fold changes from the differential expression analysis, it was revealed that several pathways were up-regulated in the basal BC that were related to the development and maintenance of structures such as the extracellular matrix, vasculartur, and blood vessels. This was unsurprising as basal cancers tend to be aggressively metastatic and would require these structures to form new tumors. Now focusing on the JAK/STAT pathway that was also up-regulated, literature confirms its place as a potential signaling cascade associated with cancer, including breast cancer [7]. The original paper was specifically focused on the JAK2/STAT3 axis as the STAT3 gene, when de-regulated, becomes highly activated and drives the expression of genes that lead to malignant behavior [4]. In our analysis, STAT3 did have a LFC of 1, but not a significant adjusted p-value. Still, the elevated values of many of the genes in the pathway, Figure 3, indicate that something was happening with this signaling cascade. In particular, the original study showed the over-activation of the JAK2/STAT3 axis lead to proliferation and metastasis *in vivo*. Therefore, it seems that this pathway warrant's more research.

The original study made the connection between the Wwox gene and the JAK2/STAT3 through several techniques. Ultimately, they claimed that Wwox impedes the ability of JAK2 to phosphorylate and then subsequently bind to STAT3, thus Wwox prevents STAT3 from becoming overactive. Given the nature of our data, we were not able to explore this result.

4.1 Limitations

As with any *in vito* RNA sequencing analysis, the use of cell lines as a sources of data may not provide the full biological picture. The human body work with multiple systems at all times in

an attempt to maintain homeostasis, so not all genes and pathways may be able to be measured in this experimental set-up. Moreover, given that Wwox expression levels may very widely in cell lines, the use of cell lines may not work to study this gene. In addition, the choice to specifically use the HBL-100 during the original study was interesting. There are at least 27 other basal breast cancer cell lines that could be used [8], so to use one that did not even originate from breast cancer could infer some bias to the results. In addition, the original study used triple negative and basal breast cancer as interchangeable terms. It is important to note, these two types of cancer are not the same thing. Thus, it may be irresponsible to use them as such.

5 Conclusions

Ultimately, we were successful in half of replicating the results from the paper of interest. Our differing results in Wwox could be due to the nature of cell lines and low number of samples. In addition, which could account for different conclusions. Our findings implicate that investigation of the JAK/STAT pathway and its therapeutic uses may prove to be a better endeavor. However, since cancer is a multi-faceted, ever-evolving adversary, it is unlikely there will ever be a magic solution.

Going forward, gene expression analysis could be completed on luminal BC that had had the Wwox gene knocked down or conversely basal BC that had the Wwox over-expressed to see the changes in other genes and pathways. Moreover, gene expression and pathway analysis could also be completed on direct tissues samples from a larger population, particularly of differing ages to measures the differences in Wwox between sub-types and ages. These studies may help to confirm the role of the this gene breast cancer. Thus, it is clear that there is an opportunity for more research with for the Wwox gene and JAK/STAT pathway.

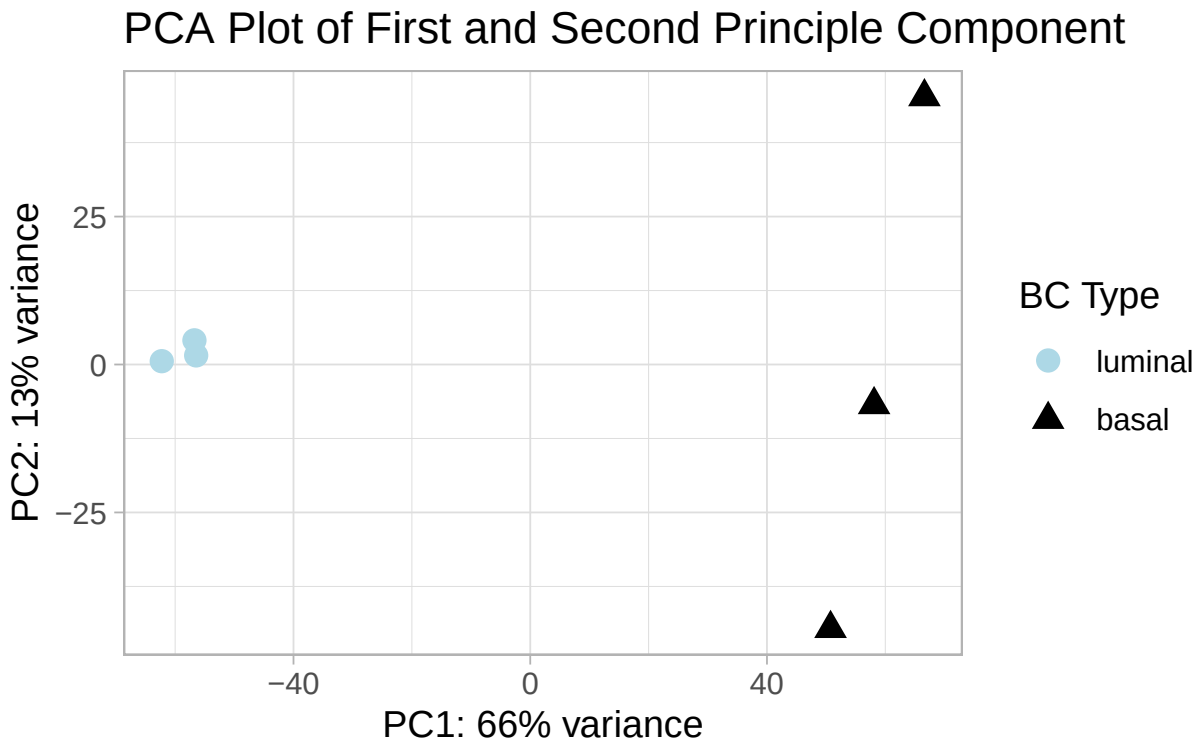


Figure 1: First and second component of PCA analysis. The first component explains 66 percent of the variation in the data and reveals a clear separation between the basal and luminal sub-types. The second component accounts for 13 percent of the variation in the data and shows the differences between the basal samples. Because the sub-types are clearly differentiated from each other, this plot suggest that there are large differences between the groups.

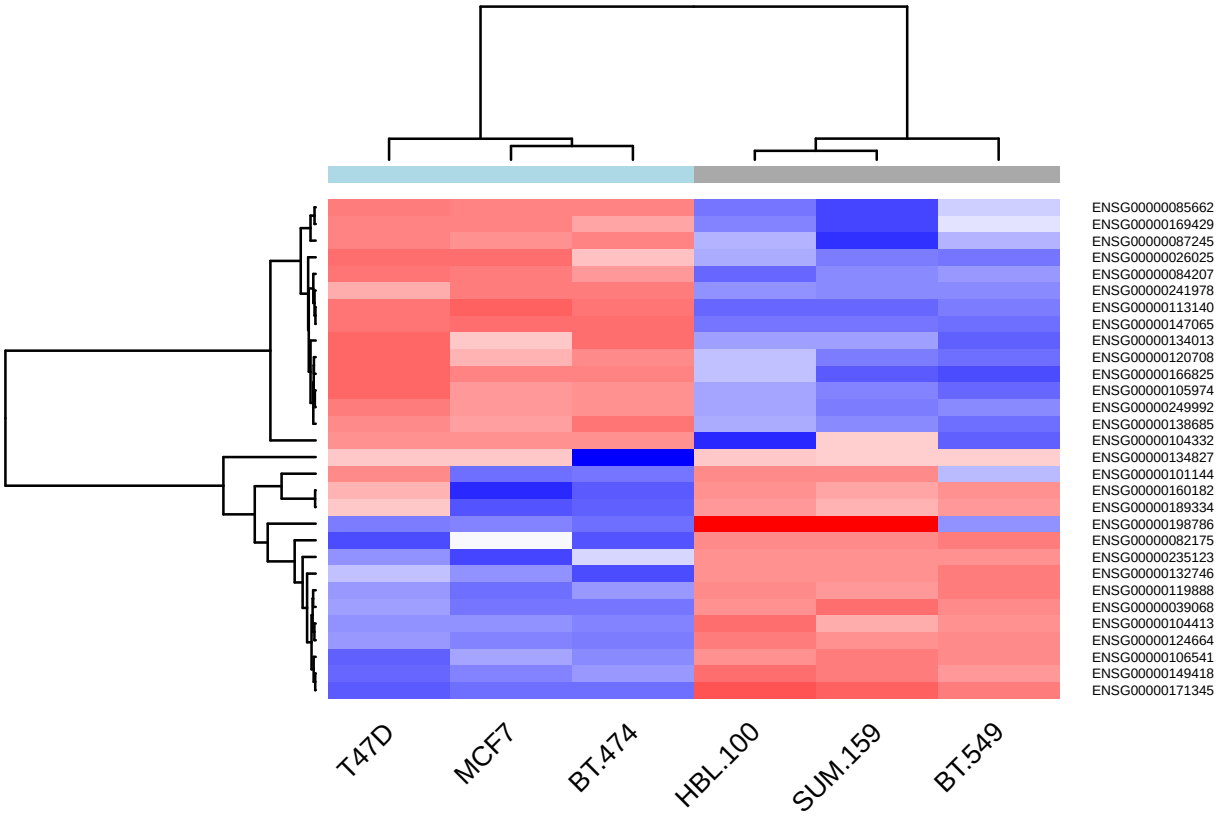


Figure 2: Hierarchical clustering plot shows clustering of the top 30 genes with the largest standard deviation and clustering of the six libraries. Color values are represented as the gene value for that sample minus the gene mean across all samples. Red values indicate that the gene value was below the mean and blue indicates that the gene value was above the mean. The libraries cluster within each sub-type. The genes cluster by those that have higher values in basal libraries versus lower values in the basal.

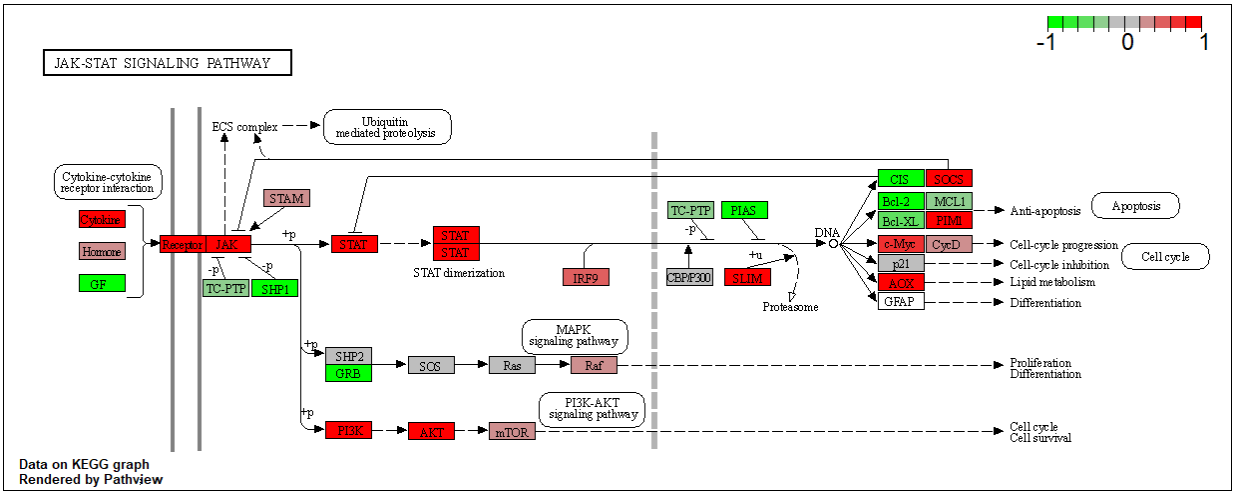


Figure 3: Map of the JAK/STAT signaling pathway as provided by KEGG. Genes that with positive LFCs (up-regulated in basal BC) are shown in red and those with negative LFCs (down-regulated in basal bc) are shown in green. These pathway conduct many signals for development and homeostasis and is related to cell cycle, cell survival, apoptosis, and other signaling pathways. Created with *Pathview* [17].

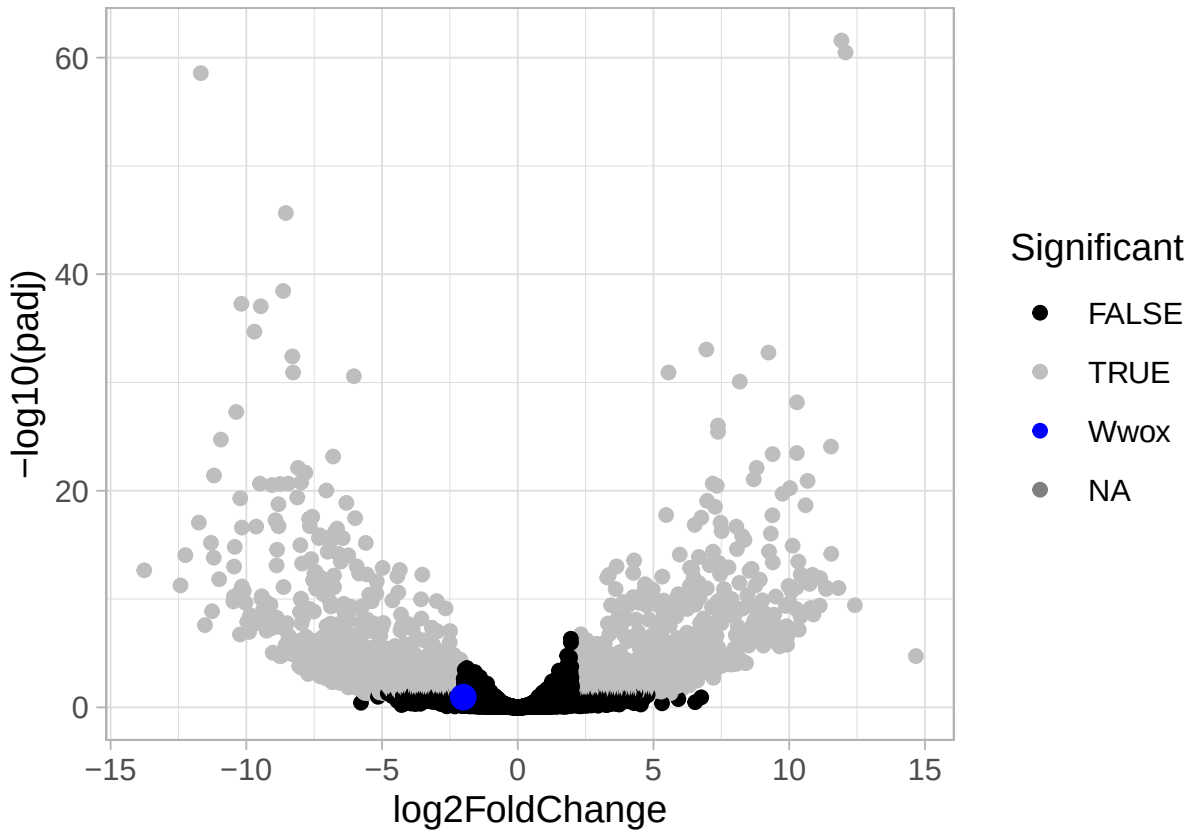


Figure 4: The volcano plot shows the LFC vs the $-log_{10}(Padj)$. Genes that were significant are shown in grey and non-significant genes are shown in black. The gene of interest, Wwox, is shown in blue, although it was not significant.The gene in the top left corner is DSCAM and the two genes in the top right are Moesin and SPARC.

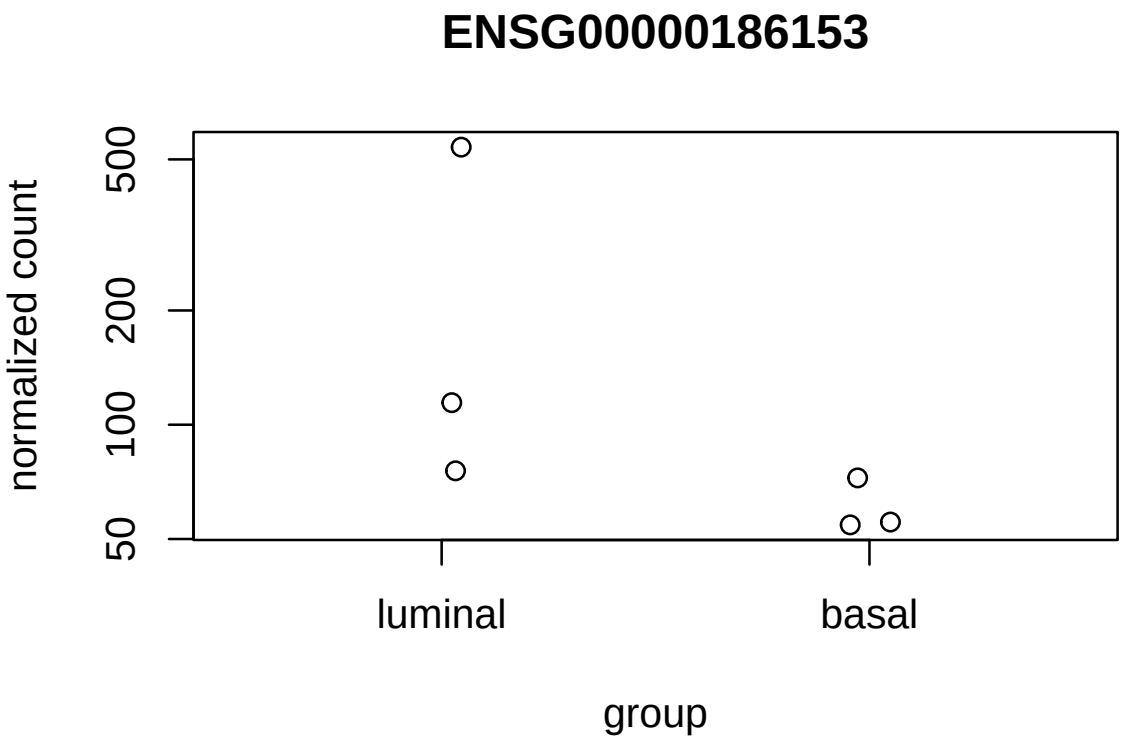


Figure 5: Normalized counts of Wwox gene by library and sub-type. On average, the basal sub-type does have lower counts than the luminal sub-type. However, this is primarily due to one luminal library, MCF-7.

Table 1: Top five GO pathways by stat mean.

Process	Direction	Mean	P-Value
Vasculature development	Greater	7.12	$1.43e^{-12}$
Blood vessel development	Greater	6.87	$7.56e^{-12}$
Extracellular matrix organization	Greater	6.78	$5.061e^{-11}$
Blood vessel morphogenesis	Greater	6.53	$7.80e^{-11}$
Angiogenesis	Greater	6.33	$2.93e^{-10}$

Table 2: Top five KEGG pathways by stat mean.

Process	Direction	Mean	P-Value
Focal adhesion	Greater	5.11	$2.76e^{-7}$
ECM-receptor interaction	Greater	4.56	$6.81e^{-06}$
Complement and coagulation cascades	Greater	3.46	$5.3e^{-04}$
Protein digestion and absorption	Greater	3.119	$2.75e^{-03}$
Jak-STAT signaling pathway	Greater	2.53	$6.03e^{-03}$

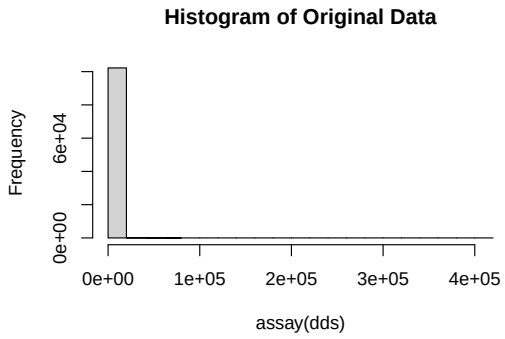
References

1. National Breast Cancer Foundation, INC. *About Breast Cancer* Accessed December 10, 2021 [Online]. https://www.nationalbreastcancer.org/about-breast-cancer/.
2. American Cancer Society. *Types of Breast Cancer* Accessed December 7, 2021 [Online]. https://www.cancer.org/cancer/breast-cancer/about/types-of-breast-cancer.html.
3. BREASTCANCER.ORG. *Molecular Subtypes of Breast Cancer* Accessed December 10, 2021 [Online]. https://www.breastcancer.org/symptoms/types/molecular-subtypes.
4. Change, R. *et al.* Loss of Wwox drive metastasis in triple-negative breast cancer by JAK2/STAT3 axis. *Nature Communications* **9** (2018).
5. Baryla, I., Styzen-Binkowska, E. & Bednarek, A. Alteration of WWOX in human cancer, a clinical view. *Experimental Biology and Medicine* **240,** 305–314 (2015).
6. Walker, S., Xiang, M. & Frank, D. STAT3 Activity and Function in Cancer: Modulation by STAT5 and miR-146b. *Cancers* **6,** 958–968 (2014).
7. Kiu, H. & Nicholson, S. E. Biology and significance of the JAK/STAT signalling pathways. *Growth Factors* **30,** 88–106.
8. Chavez, K., Garimella, S. & Lipkowitz, S. Triple Negative Breast Cancer Cell Lines: One Tool in the Search for Better Treatment of Triple Negative Breast Cancer. *Breast Disease* **32,** 35–48 (2010).
9. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2021). https://www.R-project.org/.
10. RStudio Team. *RStudio: Integrated Development Environment for R* RStudio, PBC (Boston, MA, 2021). http://www.rstudio.com/.
11. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 550 (12 2014).
12. Luo *et al.* GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10,** 161 (2009).
13. Wickham, H. *et al.* Welcome to the tidyverse. *Journal of Open Source Software* **4,** 1686 (2019).
14. Wang, C.-C. *et al.* Differential expression of moesin in breast cancers and its implication in epithelial-mesenchymal transition. *Histopathology* **61,** 78–87 (2012).
15. Basu, G. *et al.* Frequency distribution of SPARC in triple-negative breast cancer patients. *Journal of Clinical Oncology* **29.** PMID: 27957995, 37–37. eprint: https://doi.org/10.1200/jco.2011.29.27_suppl.37. https://doi.org/10.1200/jco.2011.29.27_suppl.37 (2011).
16. Jin, H., Haung, W., Tang, Q., Chen, Y. & Zhengzhi. IncRNA and breast cancer: progress from identifying mechanisms to challenges and opportunities of clinical treatment. *Molecular Therapy: Nucleic Acids,* 613–637 (2021).
17. Luo, Weijun, Brouwer & Cory. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **29,** 1830–1831 (2013).

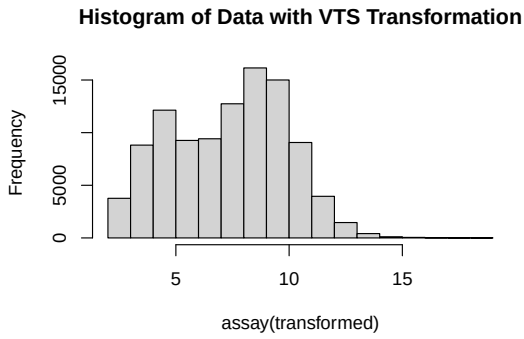
6 Appendix

R code can be found here: https://github.com/espors/RNASeq$_{B}reastCancerTypes.git$

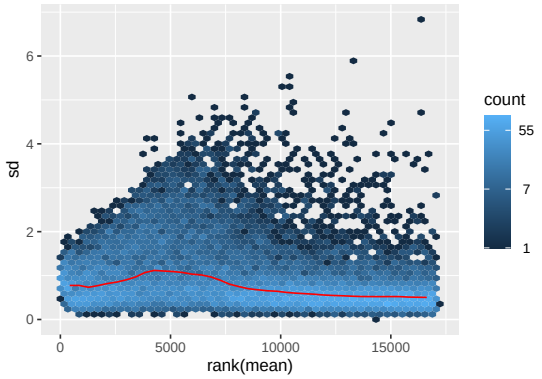
Additional images relating the analysis.



(a) Histogram of raw counts.



(b) Histogram of VST transformation.



(c) Mean-standard deviation plot of VST transformation.

Figure 6: Transformation of the raw counts. Figure 6a shows the distribution of the raw counts. Most of the values are close to zero. After transformation, Figure 6b, the data is closer to normally distributed, but the histogram is clearly bi-modal. In Figure 6c, the red line is fairly flat which indicates the transformation removed the dependence of the standard deviation on the mean. Thus, this data was determined to be acceptable for use in EDA.

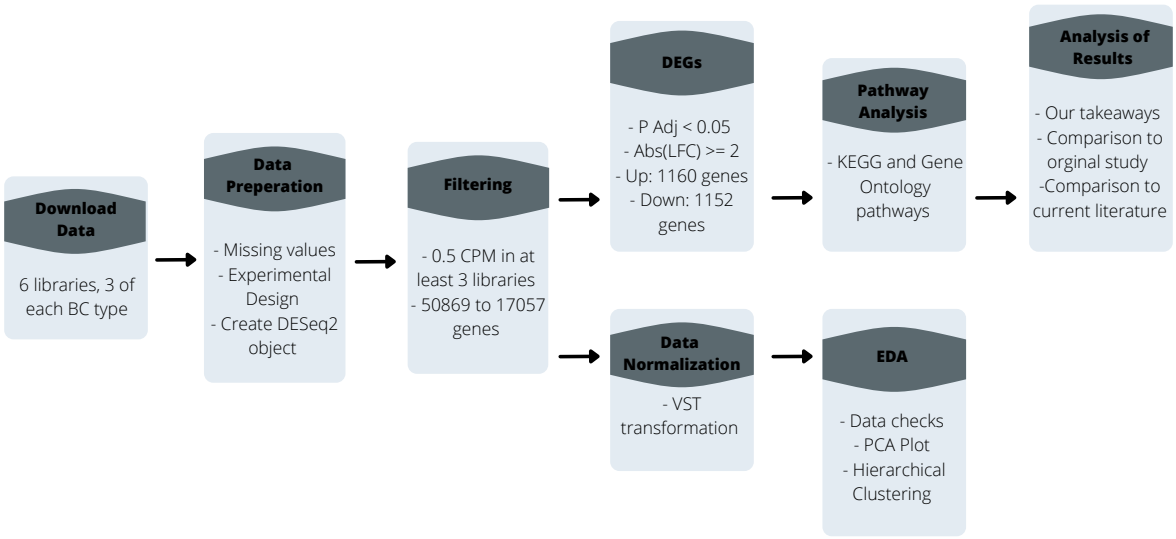


Figure 7: Complete workflow of project. This diagram describes the workflow of the project from start to end.